A computational investigation of the Sapir-Whorf hypothesis: The case of spatial relations

Christine Tseng[1] **(ctseng@caltech.edu), Alexandra Carstensen**[2] **(abc@berkeley.edu), Terry Regier**[3] **(terry.regier@berkeley.edu), Yang Xu**[3] **(yang_xu_ch@berkeley.edu)**

[1] Computation and Neural Systems, California Institute of Technology, Pasadena, CA 91125 USA
[2] Department of Psychology, University of California, Berkeley, CA 94720 USA
[3] Department of Linguistics, Cognitive Science Program, University of California, Berkeley, CA 94720 USA

Abstract

Investigations of the Sapir-Whorf hypothesis often ask whether there is a difference in the non-linguistic behavior of speakers of two languages, generally without modeling the underlying process. Such an approach leaves underexplored the relative contributions of language and universal aspects of cognition, and how those contributions differ across languages. We explore the naming and non-linguistic pile-sorting of spatial scenes across speakers of five languages via a computational model grounded in an influential proposal: that language will affect cognition when non-linguistic information is uncertain. We report two findings. First, native language plays a small but significant role in predicting spatial similarity judgments across languages, consistent with earlier findings. Second, the size of the native-language role varies systematically, such that finer-grained semantic systems appear to shape similarity judgments more than coarser-grained systems do. These findings capture the tradeoff between language-specific and universal forces in cognition, and how that tradeoff varies across languages.

Keywords: Linguistic relativity; Sapir-Whorf hypothesis; semantic universals; name strategy; categorization; spatial relations; computational models.

Introduction

Languages partition human experience into semantic categories in different ways. For example, the Mandarin Chinese spatial term *shang4* denotes a set of spatial relations that is roughly equivalent to those described by English *on* and *above* combined. Do such differences affect how speakers of different languages apprehend and think about the world?

The Sapir-Whorf hypothesis (Sapir, 1929; Whorf, 1956) is commonly framed in terms of this question. When the question is posed this way, in simple yes-or-no terms, it invites an equally simple answer: that language either does or does not influence cognition. However, empirical studies have provided conflicting answers to this question across a variety of semantic domains (Roberson, Davies, & Davidoff, 2000; Gilbert, Regier, Kay, & Ivry, 2006; Brown, Lindsey, & Guckes, 2011; Malt, Sloman, Gennari, Shi, & Wang, 1999; Munnich, Landau, & Dosher, 2001; Kranjec, Lupyan, & Chatterjee, 2014; Majid, Bowerman, Kita, Haun, & Levinson, 2004). This inconsistent pattern of results suggests that a slightly more complex formulation of the hypothesis may be warranted. Concretely, rather than asking whether language does or does not shape cognition, it may be useful to ask under which circumstances it does, to what extent, and whether that extent itself varies across languages in a systematic way, because of general principles. We pursue these questions here.

Our empirical focus is the semantic domain of topological spatial relations. Earlier investigations have revealed wide yet constrained cross-language variation in spatial semantic categories (Levinson & Meira, 2003), and have also revealed that non-linguistic cognition about such spatial relations reflects both universal forces and some influence of native language (Khetarpal, Majid, Malt, Sloman, & Regier, 2010; Carstensen et al., under revision). However, while documenting the interesting interplay of universal and linguistic forces in apparently non-linguistic spatial cognition, such earlier studies did not explore that interplay using a computational model, and did not illuminate under which circumstances language shapes cognition.

In addressing these open questions, our theoretical starting point is an influential proposal from the literature. In a classic study of language and color cognition, Kay and Kempton (1984) proposed what they called the *name strategy*: that language will affect cognition when non-linguistic aspects of cognition are ambiguous, uncertain, or otherwise ineffective. Their empirical findings were consistent with this idea. We instantiate the name strategy in a computational model, apply the model to data from the spatial domain, and explore the above open questions in terms of the theoretical framework this model provides.

In what follows, we first describe the data we consider, which are drawn from five languages: Dutch, English, Chichewa, Mandarin, and Maihiki. We then present our computational model, and show how it instantiates the name strategy. We then analyze the data through the lens of the model. To preview our results, we find: (1) that across all five languages, native language plays a small but significant role in predicting spatial similarity judgments, consistent with earlier findings, and (2) that the size of the native-language role varies systematically across languages, such that finer-grained semantic systems appear to shape similarity judgments more than coarser-grained systems do. We argue that these findings contribute to the debate over the Sapir-Whorf hypothesis by grounding the tradeoff between universal and language-specific forces in a computational model based on an independently proposed principle: the name strategy.

Data

To investigate the relation of spatial language and cognition, we compare linguistic and nonlinguistic categorization of spatial scenes across five languages: Dutch, English, Chichewa, Mandarin, and Maihiki. Maihiki is an underdocumented language of Peruvian Amazonia, presently being investigated by Lev Michael and colleagues at Berkeley. The spatial scenes were those of the Topological Relations Picture Series (TRPS) (Bowerman & Pederson, 1992). This stimulus set contains 71 spatial relational scenes, each of which depicts a spatial relationship between a figure object and a ground object. Figure 1 shows a small subset of the TRPS scenes, and the semantic categories in which these scenes fall for some of the languages we consider. It can be seen that there is considerable variation in spatial categories across these languages.

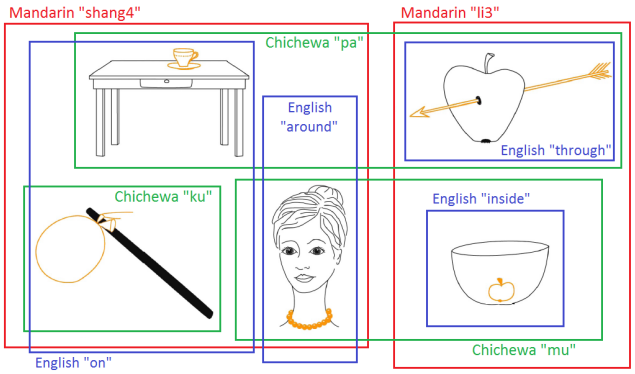

Figure 1: Cross-linguistic naming variation in the spatial domain. Each scene depicts a spatial relation between a *figure* object (in orange) and a *ground* object (in black). The scenes are grouped differently by different languages.

The data on which we rely were collected previously by Khetarpal et al. (2009, 2010, 2013), Carstensen (2011), and Carstensen et al. (under revision). Below we briefly describe the data collection procedure used in those prior studies.

Participants

A total of 47 native English speakers (24 from Khetarpal et al., 2010; 23 from Carstensen et al., under revision), 24 native Dutch speakers (Khetarpal et al., 2009), 38 native Chichewa speakers (Carstensen, 2011), 7 native Maihiki speakers (Khetarpal et al., 2013), and 17 native speakers of Mandarin Chinese (Carstensen et al., under revision) pile-sorted and then named spatial scenes. All English, Dutch, and Chichewa speakers were tested in their native languages and home countries (the United States, the Netherlands, and Malawi, respectively). Mandarin Chinese speakers were recruited on the UC Berkeley campus and tested in their native language. Speakers of Maihiki were tested in their home country of Peru, but as this is an endangered language with very few speakers, they were tested in Spanish, in which all Maihiki participants were also fluent. All participants first took part in a nonlinguistic pile-sorting task and subsequently completed a naming task; English speakers' pile-sorting data is from Khetarpal et al. (2010) but their naming data is from Carstensen et al. (under revision) for which naming instructions are more closely aligned with those of the other studies.

Pile-sorting task

Participants sorted the 71 scenes in the TRPS into piles based on the similarity of the spatial relations depicted in the scenes. Each scene showed an orange figure object positioned relative to a black ground object and participants were instructed to group the scenes into piles based on the similarity of these spatial relations, such that the relation was similar for all cards in a given pile. Participants were informed that they could make as few or as many piles as they chose, rearrange their piles as they felt necessary, and could take as much time as they wanted.

Naming task

After completing the sorting task, the same participants were asked to name the spatial relation depicted on each card. For languages other than Maihiki, labels picking out the figure and ground objects were supplied in the participant's native language and the participant filled in a blank to complete a sentence specifying the figure object's location relative to the ground object. For example, for the scene depicting a cup on a table, English-speaking participants were presented with the partial sentence "The cup is (blank) the table", and were asked to fill in the blank. Maihiki speakers were asked to produce full sentences, supplying names for the figure and ground objects and describing the spatial relation between them; verbal clarification of the scenes was given in Spanish when necessary. In keeping with earlier work, the labels produced in the naming task were sanitized to collapse over responses that differed in components without spatial meaning (e.g. variations in verb tense).

Treatment of the data

We aggregated these data into separate language-specific and universal components, for use in our computational analyses below. For each language l, we constructed a 71×71 *co-naming* matrix L^l, such that entry L^l_{ij} of that matrix contained the proportion of speakers of language l who supplied the same spatial term for scenes i and j. For instance, if 12 out of 16 speakers of language l supplied the same spatial term in l for scenes i and j, then $L^l_{ij} = 12/16 = 0.75$. Thus the L^l matrix summarizes which scenes tended to receive the same name vs. different names in language l. We analogously constructed, for each language l, a 71×71 *co-sorting* matrix S^l, such that entry S^l_{ij} of that matrix contains the proportion of speakers of language l who sorted scenes i and j into the same pile. Finally, we approximated a universal similarity space, U, by taking the average over the S^l matrices across languages l. Earlier studies (Khetarpal et al., 2010; Carstensen et al., under revision) found that spatial sorting patterns were

broadly similar across languages,[1] although they did reflect the sorter's native language to some extent. Thus, U is an attempt to retain what is common and discard what is different about pile-sorting behavior across languages.

Computational formulation

The core principle behind our analyses is the name strategy of Kay and Kempton (1984). Figure 2 illustrates this principle in the context of a simple pile-sorting scenario. Suppose there are three stimuli (in our case three spatial scenes) here labeled 1, 2, and 3, and suppose that the task is to sort stimulus 3 into one of two existing piles, which presently contain stimuli 1 and 2 respectively. Assume further that the three stimuli are equally distant from each other in a universal similarity space, so that it is entirely ambiguous on that basis which pile stimulus 3 should be sorted into. Finally, assume that the speaker's native language partitions these stimuli into two categories $N1$ and $N2$, where stimuli 1 and 3 are co-named under $N1$, and stimulus 2 is named under $N2$. The name strategy holds that in such cases, where universal non-linguistic structure yields ambiguity or is otherwise ineffective, linguistic category structure may provide additional information to resolve the issue. In this case, stimuli 3 and 1 are co-named, which should encourage stimulus 3 to be sorted into pile 1, tipping the balance in that direction. The bars in the bottom panel of Figure 2 contrast the choice probabilities over the two piles for a hypothetical model that relies only on a universal similarity space, with those for a model that relies only on the category structure of the language.

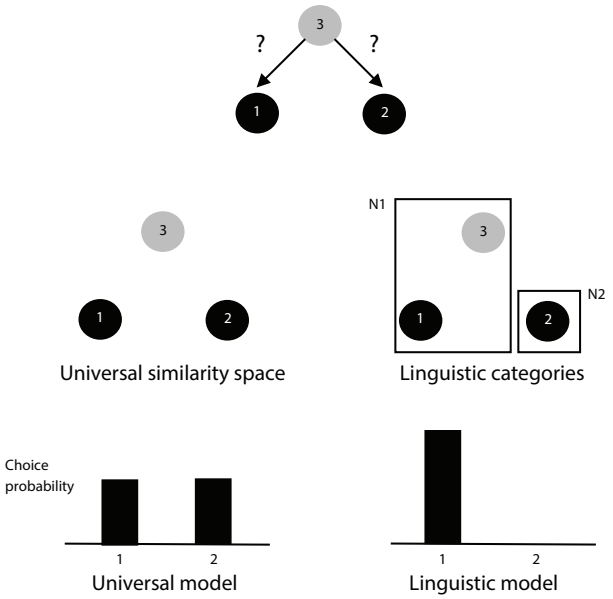

Figure 2: Illustration of the name strategy.

To capture the idea of the name strategy, we formulate models that can be used to predict an individual's pile-sorting behavior. These predictions are based on universal similarity structure U, and on language-specific naming information L^l, both as specified above, where l is the native language of the individual in question. We instantiate the name strategy through a *residual predictive analysis*: we first note how much of an individual's sorting behavior can be predicted by U, and then determine how much of that individual's as-yet-unexplained (residual) sorting behavior can be predicted by L^l, beyond what U can predict. This two-step procedure requires that we specify a *universal model* based on U that we employ in the first step, and a *language-specific model* based on L^l for the residual analysis, i.e. the items left unexplained by U. This procedure provides a way to quantify the relative contributions of universal and language-specific forces, and to assess the degree to which those relative contributions may vary across languages.

For each individual, we know the number of piles that individual produced, and which scenes were sorted into each pile. We sought to recapitulate the sorting process of that individual. Concretely, for each query scene, we sought to predict which pile that scene was sorted into, using a leave-one-out procedure. That is, for each scene i, we held out pile membership for i and sought to predict pile membership for i, based on pile membership for all other scenes $j \neq i$.[2] We cast this prediction in probabilistic terms, and predicted that this individual would place spatial scene i in that pile c (for non-linguistic category) that yielded the highest posterior probability $p(c|i)$:

$$p(c|i) \propto p(i|c)p(c) \propto f(i|c) \tag{1}$$

Here, we assume the individual has no preference for any pile to begin with, and we therefore place a uniform prior on c. In determining the likelihood $f(i|c)$, we considered three possible strategies on which individuals might rely in sorting scenes into piles, shown in Equation 2 below. These are: (a) a fixed clustering strategy: sort by average similarity between a query scene i and all existing scenes j in a given pile c; (b) a fixed nearest-neighbor or chaining strategy: sort based on maximum similarity between a query scene i and any of the existing scenes j in a given pile c; and (c) a hybrid strategy that varies on a trial-by-trial basis: choose between the clustering and chaining strategies according to which yields higher likelihood. We formalize these strategies by specifying the likelihood function as follows:

$$f(i|c) = \begin{cases} \frac{1}{|c|}\sum_{j\in c} sim(i,j), & clustering \\ \max_{j\in c} sim(i,j), & chaining \\ \max(f(i|c)_{clustering}, f(i|c)_{chaining}), & hybrid \end{cases} \tag{2}$$

[1] We verified the universal tendencies in pile sorting by correlating the S matrices (upper triangular parts due to symmetry) between each pair of languages. We confirmed that pile sorting is largely similar across speakers of different languages reflected in mean Pearson's $r = 0.98$ ($SD = 0.01$) among the pairwise correlations.

[2] Ideally, we would like to predict scene-pile assignments in the sequence in which they occurred during the experiment. However, only the end-state of the pile-sort was recorded, not the sequence that led to it, so we used the leave-one-out procedure which is unaffected by the sorting sequence.

For each scene assignment, we chose the strategy that yielded the highest predictive accuracy. To distinguish between the universal and language-specific models, we took $sim(i,j) = U_{ij}$ for the universal model, and $sim(i,j) = L^l_{ij}$ for the linguistic model, where l is the individual's native language and U and L^l are as defined above.

To model a given individual's behavior, we first conducted a predictive analysis using the universal model, and noted which of that individual's scene assignments were predicted correctly and which incorrectly. We then conducted a residual predictive analysis using the language-specific model, on those scenes that were incorrectly predicted by the universal model, and again noted which scenes were predicted correctly and which incorrectly.

Analyses and results

We conducted such analyses for each speaker of each language. We then examined the results of those analyses with a view to answering the open questions posed at the beginning of this paper. We did so in three sets of followup analyses, which we present below.

Relative contributions of universal and language-specific forces

One open question is the magnitude of the relative contribution of universal and language-specific forces to allegedly non-linguistic tasks such as pile-sorting, when assessed using the method outlined above.

Figure 3 summarizes the results of the residual predictive analyses just described. For each language, the black bar shows the accuracy of the universal model in predicting scene-pile assignments, averaged across speakers of that language. The white bar stacked on top of it shows the accuracy of the language-specific model in predicting residual scene-pile assignments, i.e. those that the universal model failed to predict, again averaged across speakers of that language.[3] To establish a baseline in this language-specific residual predictive task, we considered chance predictions from a model that chooses scene-pile assignments randomly from among the available piles. In this case, the chance-level accuracy for each individual is $\frac{1}{k}$ where k is the number of piles that individual generated during pile sorting, and the dashed horizontal line for each language shows the chance level of residual predictive accuracy for that language, averaged across speakers of the language.

Overall, the universal model accounts for a substantial proportion of the pile-sort data for each language, suggesting strong universal tendencies in people's similarity judgments about spatial relations. At the same time, for each language, the language-specific model predicts residual scene-pile assignments at rates above chance. To assess whether this effect of language is significant at the level of individual speakers, we examined the proportion of individuals

[3]Similar results were also obtained when, for each speaker, we left pile-sort data from that speaker's native language out of the universal similarity matrix U on which we base predictions.

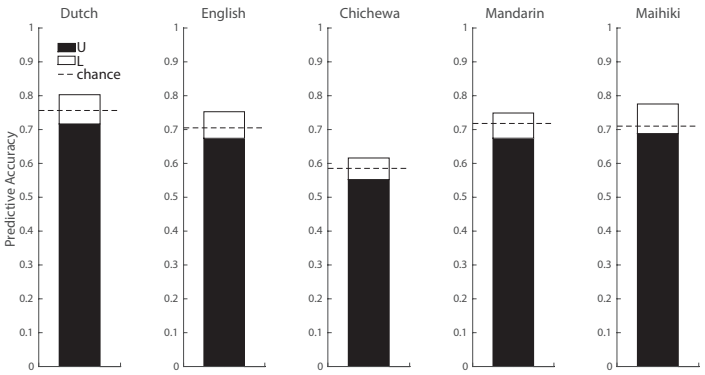

Figure 3: Summary of results of residual predictive analyses.

for whom the language-specific model exhibits above-chance residual predictive accuracy. We found that this proportion is high across speakers of all the languages we considered (Chichewa: 30/38; Dutch: 21/24; English: 20/23; Maihiki: 6/7; Mandarin: 15/17; Bonferroni-corrected $p < 0.05$ for each language except for Maihiki (uncorrected $p = 0.05$) under binomial tests assuming 0.5 probability of success per speaker). Taken together, these findings provide evidence in support of the Sapir-Whorf hypothesis in this domain, and contextualize it relative to what can be accounted for by universal forces alone. It should be emphasized that this is a rather conservative test for an effect of language, in that any scene-pile assignment that would be correctly predicted by both the universal model and the language-specific model will be credited here to the universal model, as it was run first, in keeping with our instantiation of the name strategy. Thus, it may be safest to think of these residual predictive accuracies as providing a lower bound on the size of the language-specific contribution.

Native language compared with other languages

Support for the Sapir-Whorf hypothesis would be strengthened if we had evidence that an individual's native language outperformed other languages in residual predictive accuracy. For example, is residual Maihiki pile-sorting better predicted by Maihki naming than it is by Chichewa naming, or Mandarin, or some other language? This pattern would be expected if the sorter's native language is in fact being called upon during the putatively non-linguistic sorting process. We turn next to consider this question.

To test this, we re-ran the residual predictive analyses described above, but for each individual, instead of using that individual's native-language naming information (in the form of L^l for native language l), we used naming information from each other language (i.e. L^k for each language $k \neq l$). For example, to predict residual pile-sorting data for Dutch speakers, we used four different linguistic models based on naming from each of the four alternative languages other than Dutch, while keeping all other procedures unchanged. For this and remaining analyses, we focused on individuals that exhibited

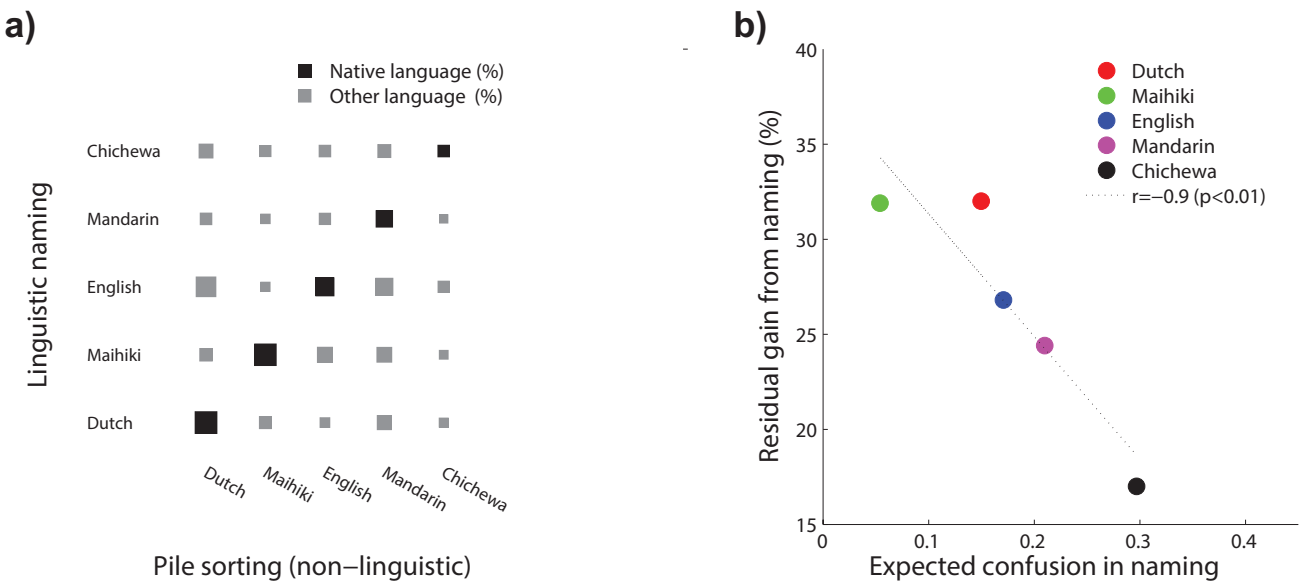

Figure 4: Summary of results of cross-language comparisons. (a) Within-language and cross-language predictive accuracies on residual scene-pair assignments. Diagonal elements (black squares) reflect predictive accuracies from native languages. Off-diagonal elements (gray squares) reflect predictive accuracies from other languages. Square size is proportional to percentage gain in predictive accuracy in the residual analysis. (b) Relationship between naming confusion (semantic coarse-grainedness; horizontal axis) and percentage gain in residual prediction from each language (vertical axis).

above-chance accuracies in residual prediction.

Figure 4(a) presents the results of this cross-language predictive analysis. The plot summarizes how well different linguistic naming matrices L (by row) predict residual scene-pile assignments in pile-sorting by speakers of different languages (by column). Square size is proportional to percentage gain in predictive accuracy in the residual analysis. Diagonal elements (in black) represent performance given native-language information, and the four off-diagonal elements (in gray) within each column represent performance given non-native-language information, on the same pile-sort data. In almost all cases, native-language models outperformed the non-native-language models in predicting residual pile-sorting data by speakers of that native language (19 out of 20 pairwise comparisons; the exception is that English naming predicts sorting data from Mandarin speakers slightly better than Mandarin naming itself). This finding supports the suggestion that native language was recruited in pile-sorting. We note also that the percentage gain in the residual prediction differs across languages. In Figure 4(a), the diagonal elements are sorted by accuracy, and it can be seen that Dutch predicts its speakers' residual pile-sort data the best, and Chichewa predicts its speakers' residual pile-sort data the worst. In our final analysis, we asked whether this variation in cross-language predictive accuracy is systematically linked to the nature of the semantic systems involved.

Semantic grain and linguistic relativity

Our final analysis builds on earlier explorations of *semantic grain* in sorting and naming (Khetarpal et al., 2010; Carstensen et al., under revision). A natural possibility is that fine-grained semantic systems may have a greater effect on non-linguistic spatial similarity judgments (reflected in higher residual predictive accuracy) than coarse-grained systems do. The rationale is that a fine-grained semantic system offers more opportunities to resolve ambiguous cases. This proposal follows from the name strategy, where Kay and Kempton (1984) found that the degree of linguistic effect on how speakers resolve ambiguity in color judgements (e.g. distinguishing colors near the blue-green boundary) depends on the fine-grainedness of color naming systems. Similarly, in the case of pile-sorting of spatial scenes, we expect scene pairs that cannot be resolved linguistically by a coarse-grained system because they fall in the same category in that system, can be resolved linguistically by a fine-grained system because they fall in different categories in that system.

To test this prediction, we examined the relationship between native-language residual predictive accuracy (from the previous analyses) and the semantic grain of each language. We assessed semantic grain for each language l by measuring the expected amount of *naming confusion*, or co-naming, in l's co-naming matrix L^l, averaging together the co-naming proportions across all unique pairs of scenes i, j:

$$nc(l) = \text{average}_{(i,j)} L^l_{i,j} \tag{3}$$

Here, $nc(l)$ measures the extent to which different stimuli tend to receive the same name and thus be linguistically indistinguishable in language l. Coarse semantic grain corresponds to a high value for $nc(l)$ (because different scenes i, j will often receive the same names in a coarse-grained system,

and thus have high co-naming values $L^l_{i,j}$), and fine semantic grain corresponds to a low value for $nc(l)$.

Figure 4(b) shows that there is a strong negative correlation (Pearson's $r = -0.9$; $p < 0.01$ from a permutation test with 10,000 samples) between expected naming confusion (coarse-grainedness) for a language, and residual gain in predictive accuracy (size of the Sapir-Whorf effect) for that language.[4] Dutch and Maihiki–the two most fine-grained systems in our data–yield the highest predictive accuracies in the residual analysis. In contrast, Chichewa–which has a comparatively coarse-grained spatial semantic system–yields the lowest residual predictive accuracy. We have considered only five languages here, and future work can usefully examine the robustness of this finding with a larger set of languages. Nevertheless, these findings provide initial support for the prediction that semantically finer-grained languages will tend to have a greater effect on non-linguistic judgments than coarser-grained languages will.

Conclusion

We have presented two main findings: (1) Language appears to play a small but significant role in shaping spatial similarity judgments, in line with earlier studies, and (2) the extent of this linguistic effect varies as a function of the semantic grain of one's native language. We have arrived at these findings by pursuing a proposal from the literature, Kay and Kempton's (1984) name strategy, and by instantiating that proposal in a computational analysis. Similar ideas appear elsewhere in the literature (e.g. Vong et al., 2015), and we hope that our work will encourage further formal analyses of the link between cross-language semantic structures and cognition.

Acknowledgments

We thank Lev Michael and Grace Neveu for access to the Maihiki data, and Asifa Majid for access to the Dutch data. This project was funded by NSF award SBE-1041707, the Spatial Intelligence and Learning Center (SILC), and NSF Graduate Research Fellowship grant DGE-1106400 (to AC).

References

Bowerman, M., & Pederson, E. (1992). Topological relations picture series. In *Space stimuli kit 1.2* (p. 51). Nijmegen: Max Planck Institute for Psycholinguistics.

Brown, A., Lindsey, D., & Guckes, K. (2011). Color names, color categories, and color-cued visual search: Sometimes, color perception is not categorical. *Journal of Vision*, *11*.

Carstensen, A. (2011). *Universals and variation in spatial language and cognition: Evidence from Chichewa.* Undergraduate thesis, University of California, Berkeley.

Carstensen, A., Khetarpal, N., Majid, A., Malt, B., Sloman, S., & Regier, T. (under revision). Cross-language universals and variation in cognition: The cases of space and artifacts.

Gilbert, A., Regier, T., Kay, P., & Ivry, R. (2006). Whorf hypothesis is supported in the right visual field but not the left. *Proceedings of the National Academy of Sciences*, *103*, 489-494.

Kay, P., & Kempton, W. (1984). What is the Sapir-Whorf hypothesis? *American Anthropologist*, *86*, 65-79.

Khetarpal, N., Majid, A., Malt, B., Sloman, S., & Regier, T. (2010). Similarity judgments reflect both language and cross-language tendencies: Evidence from two semantic domains. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*.

Khetarpal, N., Majid, A., & Regier, T. (2009). Spatial terms reflect near-optimal spatial categories. *Proceedings of the 31st Annual Meeting of the Cognitive Science Society*.

Khetarpal, N., Neveu, G., Majid, A., Michael, L., & Regier, T. (2013). Spatial terms across languages support near-optimal communication: Evidence from Peruvian Amazonia, and computational analyses. *Proceedings of the 35th Annual Meeting of the Cognitive Science Society*.

Kranjec, A., Lupyan, G., & Chatterjee, A. (2014). Categorical biases in perceiving spatial relations. *PLoS ONE*, *9*, e98604.

Levinson, S. C., & Meira, S. (2003). Natural concepts in the spatial topological domain–adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, *79*, 485-516.

Majid, A., Bowerman, M., Kita, S., Haun, D., & Levinson, S. (2004). Can language restructure cognition? The case for space. *Trends in Cognitive Sciences*, *8*(3), 108-114.

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*(2), 230–262.

Munnich, E., Landau, B., & Dosher, B. (2001). Spatial language and spatial representation: A cross-linguistic comparison. *Cognition*, *81*, 171-207.

Roberson, D., Davies, I., & Davidoff, J. (2000). Color categories are not universal: Replications and new evidence from a stone-age culture. *Journal of Experimental Psychology: General*, *129*(3), 369-398.

Sapir, E. (1929). The status of linguistics as a science. *Language*, 207–214.

Vong, W. K., Navarro, D. J., & Perfors, A. (2015). The helpfulness of category labels in semi-supervised learning depends on category structure. *Psychonomic Bulletin & Review*, 1–9.

Whorf, B. (1956). The relation of habitual thought and behavior to language. In J. Carroll (Ed.), *Language, thought, and reality* (p. 134-159). Cambridge, Massachusetts: The MIT Press.

[4] We replicated this finding using a discrete measure for semantic grain based on height following Khetarpal et al., (2010), where we found a strong negative correlation between mean coarseness in naming for a language and residual gain (Pearson's $r = -0.92$).